UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/647,203	08/21/2003	Alexander Franz	24207-10274	1475

62296 7590 02/22/2008

GOOGLE / FENWICK
SILICON VALLEY CENTER
801 CALIFORNIA ST.
MOUNTAIN VIEW, CA 94041

EXAMINER	
SHAH, PARAS D	

ART UNIT	PAPER NUMBER
2626	

MAIL DATE	DELIVERY MODE
02/22/2008	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary	Application No.	Applicant(s)	
	10/647,203	FRANZ ET AL.	
	Examiner	Art Unit	
	Paras Shah	2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *12/26/2007*.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1,3-6,8-14,16-25 and 27-54* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1,3,4,8,9,11-14,16-18,20-24,27-29 and 31-54* is/are rejected.

7)☒ Claim(s) *5,10,19,25 and 30* is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☐ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____.
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

1. This communication is in response to the RCE filed on 12/26/2007. Claims 1, 3-6, 8-14, 16-25, 27-54 remain pending and have been examined. The Applicants' amendment and remarks have been carefully considered but they do not place the application in condition for allowance.

2. All previous objections and rejections directed to the Applicant's disclosure and claims not discussed in this Office Action have been withdrawn by the Examiner.

Response to Arguments

3. Applicant's arguments (pages 14-17) filed on 12/26/2007 with regard to 1, 3-6, 8-14, 16-25, 27-36 have been fully considered and are moot in view of new grounds for rejection.

Response to Amendment

4. Applicants' amendments filed on 12/26/2007 have been fully considered. The newly amended limitations in claims 1, 3-6, 8-14, 16-25, and 27-54 necessitate new grounds of rejection.

Claim Rejections - 35 USC § 103

5. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically teach or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention

was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

6. Claims 1, 3, 6, 8, 11-13, 20-24, and 31-54 are rejected under 35 U.S.C. 103(a) as being unpatentable over Su *et al.* (In *Proceedings of the 32nd Annual Meeting on Association For Computational Linguistics* 1994) in view of Frantzi *et al.* ("Extracting Nested Collocations").

As to claims 1, 6, and 12, Su *et al.* teaches a system for finding compound words in a text corpus comprising:

a vocabulary (see page 244, 2nd full paragraph, sect. Simulation, (1st paragraph), line 5-8) comprising of tokens (see page 244, Table 1) from a text corpus (see page 243, left column, 2nd paragraph, line 6)

a compound finder executable to iteratively finding compounds having a plurality of length within the text corpus, and rebuild at least part of (see page 245, right column, "Simulation," 1st paragraph, compound list is modified or rebuild after a new compound word is detected. The compounds having plurality of lengths is obvious in document being studied.) the vocabulary based on the identified compounds having plurality of lengths (see each compound comprising a plurality of tokens, comprising: (page 244, left column, 1st paragraph, line 10) (e.g. It should be noted that windowing the corpus in sizes of 2 and 3 over the text corpus can be interpreted as a form of iteration when finding compounds of these various lengths), the compound finder comprising,

n-gram counter (see page 244, left column, 1st paragraph, lines 3-4) executable to evaluate a frequency of occurrence (n-gram counter) (see

page 244, left column, 1st paragraph, lines 3-4) for one or more n-grams (see page 243, left column, 3rd paragraph, lines 1-5) and

a likelihood evaluator executable to:

determine a likelihood of collocation for one or more of the n-grams having the same length compounds (see page 243, right column, line 8),

add a subset of n-grams that satisfy at least one criterion evaluated responsive to the likelihood of collocation (see page 245, right column, 2nd paragraph, line 7) (e.g. A subset can be 0 or more and this is done by the reference by adding those compounds greater than 1.);

and rebuild at least part of the vocabulary based on the added subset of n-grams (see page 245, right column, "Simulation", 1st paragraph, vocabulary is rebuild by the adding of more compound words and the sorting).

Su *et al.* does not specifically teach the use of an iterator for selecting n-grams having a length that is less than the selected n-gram. It should be noted that Su *et al.* does suggest using window sizes of two or three for n-gram determination (see page 243, left column, 1st paragraph) (e.g. It as noted in the Applicant's arguments that Su teaches away from the iteration. The Examiner traverses this argument. Su *et al.* makes no mention that the implementation for determining the compounds of lengths 2 and 3 are not done using iteration.

Thus, one skilled in the art would have implemented iteration as is well known method for implementing the methods of Su in a systematic manner. Further, it should also be noted that the arguments made with respect to a two cluster classifier, it does not have to do with it being a bigram or a trigram but rather determining if it is a compound/non-compound (see page 243, right column, lines 1-7).

However, Frantzi *et al.* does teach the use of an iterator executable to select n-grams having a same length that is less than a length of n-grams selected during a previous iteration (page 43, right column, "The algorithm ...", 2nd full paragraph, code underneath and page 44, entire left column-right column, numbered item 5) (e.g. From the cited reference it is seen that the n-gram starts from some maximum limit and then proceeds to a lower order n-gram. The n-gram is decremented and takes into account the frequency of occurrence in order to determine a candidate collocation by the determination of a C value.)

It would have been obvious to one of ordinary skilled in the art at the time the invention was made to have modified the finding of compounds words in a corpus as taught by Su *et al.* with the backward iteration as taught by Frantzi *et al.* The motivation to have combined the references involves the ability to systematically determine the likelihood of collocation and extract the unextracted collocations that occur (see Abstract) and thus making the process automatic.

As to claims 3 and 8, Su *et al.* in view of Frantzi *et al.* teach all of the limitations as claim 1 above.

Furthermore Su *et al.* teaches a system where only some of the subset of n-grams that have a high likelihood are added as compounds to the vocabulary (page 245, right column, 2nd paragraph, line 6-8) (e.g. It should be noted that the selection of those compounds, which have a high likelihood will be chosen if the value is greater than 0, otherwise it will not be included).

As to claim 11, , Su *et al.* in view of Frantzi *et al.* teach all of the limitations as in claim 6 above.

However, Su *et al.* in view of Frantzi *et al.* do not specifically teach the use of a computer for compound extraction. Su *et al.* does mention simulation for compound extraction (see Su *et al.* page 245, right column, 2nd paragraph). Hence, it is obvious to one of ordinary skilled in the art to have used a computer to execute the simulation from code. The motivation to include a computer-storage medium is for use in machine translation (see Su *et al.* page 243, left column, 1st paragraph, line 27).

As to claims 13, 24, and 36 Su *et al.* teaches a system for identifying compounds through iterative analysis comprising:

a compound finder executable to iteratively finding compounds having a plurality of length within the text corpus, and rebuild at least part of (see page

245, right column, "Simulation," 1st paragraph, compound list is modified or rebuild after a new compound word is detected. The compounds having plurality of lengths is obvious in document being studied.) the vocabulary based on the identified compounds having plurality of lengths (see each compound comprising a plurality of tokens, comprising: (page 244, left column, 1st paragraph, line 10) (e.g. It should be noted that windowing the corpus in sizes of 2 and 3 over the text corpus can be interpreted as a form of iteration when finding compounds of these various lengths), the compound finder comprising,

n-gram counter executable to: (see Su *et al.* page 244, left column, 1st paragraph, lines 3-4)

determine a number of occurrences of one or more n-grams (e.g. The maximum number of tokens depends on the iteration value or step) the number of tokens up to the limit for iteration (see Su *et al.* page 243, left column, 2nd paragraph, line 3 and line 10) (e.g. A limit is pre-specified by the reference), which are at least in part provided in a vocabulary for the text corpus (see page 244, Table 1) from a text corpus(see page 24, 2nd full paragraph, sect. Simulation, (1st paragraph), line 5-8).

a likelihood evaluator executable to: (see Su *et al.* page 243, right column, line 8).

determine a measure of association between tokens (see Su *et al.* page 243, right column, lines 20-23) and , which adds the compound words having a high likelihood to the vocabulary (see Su

et al. page 245, right column, 2nd paragraph, line 7). Further, the adjustment of the limit can also be interpreted as the change in the n value of an n-gram. Thus, the change of limit from n=2 to n=3, will change the number of tokens per compound (see Su *et al.* page 243, left column, 2nd paragraph, lines 9-10). However, Su *et al.* does not specifically teach the use of a stored limit of the number of tokens per compound and the use of a vocabulary. It would have been obvious to one of ordinary skilled in the art to have included a predetermined limit on the number of token per compound. The motivation to modify the compound extraction by Su *et al.* by the inclusion of a stored limit is to acquire the compounds of interest to the user (see Su *et al.* page 243, 2nd paragraph, line 6) (e.g. The reference uses n-grams of n=2, and n=3).

add each identified n-gram with a sufficient measure of association to the vocabulary as a compound token (see page 245, right column, 2nd paragraph, line 7) (e.g. A subset can be 0 or more and this is done by the reference by adding those compounds greater than 1.);

and rebuild at least part of the vocabulary based on the added subset of n-grams (see page 245, right column, "Simulation", 1st paragraph, vocabulary is rebuild by the adding of more compound words and the sorting).

Su *et al.* does not specifically teach the use of an iterator for selecting n-grams having a length that is less than the selected n-gram.

However, Frantzi *et al.* does teach the use of an iterator executable to initially specify a limit (see page 44, right column, sect. 4, 1st and 2nd paragraph, limit of 10 is used.) on a number of tokens per compound for an iteration and decreasing the limit for a subsequent iteration (page 43, right column, "The algorithm ...", 2nd full paragraph, code underneath and page 44, entire left column-right column, numbered item 5) (e.g. From the cited reference it is seen that the n-gram starts from some maximum limit and then proceeds to a lower order n-gram. The n-gram is decremented and takes into account the frequency of occurrence in order to determine a candidate collocation by the determination of a C value.)

> identify at least one n-gram comprising a number of tokens equal to the limit for the iteration (see page 44, right column, sect. 4, 1st paragraph, n-grams of size 10 are extracted.) based on the number of occurrences (See page 44, right column, 2nd paragraph, those with a specific number of occurrences are used.);

It would have been obvious to one of ordinary skilled in the art at the time the invention was made to have modified the finding of compounds words in a corpus as taught by Su *et al.* with the backward iteration as taught by Frantzi *et al.* The motivation to have combined the references involves the ability to

systematically determine the likelihood of collocation and extract the unextracted collocations that occur (see Abstract) and thus making the process automatic.

As to claims 20-21 and 31-32, Su *et al.* and Frantzi *et al.* teach all of the limitations as claim 13 above.

Furthermore Su *et al.* teaches an initial vocabulary (see page 24, 2nd full paragraph, sect. Simulation, (1st paragraph), line 5-8) where token are extracted from a text corpus (see page 243, left column, 2nd paragraph, lines 6-9) through morphological analysis (e.g. It should be noted that morphological analysis and parsing is similar).

As to claims 22 and 33, Su *et al.* and Frantzi *et al.* teach all of the limitations as claim 13 above.

Furthermore Su *et al.* teaches a filter determining the number of occurrences of one or more n-grams within the text corpus for unique n-grams (see page 243, left column, 1st paragraph, line 3 and lines 7-9) (e.g. It should be noted that the use of the relative frequency is a measure for compound extraction and can thus be interpreted as a filtering means when the compound filtering is done) (see page 243, left column, 1st paragraph, lines1-5).

As to claims 23 and 34, Su *et al.* and Frantzi *et al.* teach all of the limitations as claim 13 above.

Furthermore Su *et al.* teaches a system where the text corpus comprises of documents comprising one of a news message and text (see Su *et al.* abstract).

As to claim 35, Su *et al.* in view of Frantzi *et al.* teach do not specifically teach the use of a computer for compound extraction. Su *et al.* does mention simulation for compound extraction (see Su *et al.* page 245, right column, 2nd paragraph). Hence, it is obvious to one of ordinary skilled in the art to have used a computer to execute the simulation from code. The motivation to include a computer-storage medium is for use in machine translation (see Su *et al.* page 243, left column, 1st paragraph, line 27).

As to claims 37, 40, 43, 46, 49, and 52, Su *et al.* in view of Frantzi teaches all of the limitations as in claims 1, 6, 12, 13, 24, and 36, above.

Furthermore, Su *et al.* teaches where in the added subset of n-grams satisfy a criterion having a highest measure of collocation (see page 243, right column, lines 16-23, when lambda is greater than 1, a high likelihood of collocation exists and added to compound cluster. And see page 245, right column, "Simulation", 1st paragraph, if lambda is greater than zero then the n-gram is included in the compound list.)

As to claims 38, 41, 45, 47, 50, 53, Su *et al.* in view of Frantzi teaches all of the limitations as in claims 1, 6, 12, 13, 24, and 36, above.

Furthermore, Su *et al.* teaches wherein a number of n-grams in the added subset of n-gram is equal to a defined number which specifies a maximum number of n-grams having a highest likelihood of collocation to be added (see page 245, right column, "Simulation, 1st paragraph) (e.g. In the cited section, a compound is added depending on the likelihood of collocation exceeding a threshold. Further, 1 compound is added, which was interpreted to be defined since the subset can be either interpreted to be 0 or more words. Hence, if one collocation is evaluated and determined to be compound that is a defined maximum of 1, but can also be zero that can be added.).

As to claims 39, 42, 45, 48, 51, 54, Su *et al.* in view of Frantzi teaches all of the limitations as in claims 1, 6, 12, 13, 24, and 36, above.

Furthermore, Su *et al.* teaches wherein the likelihood of collocation for each n-gram of the added subset of n-grams satisfy a criterion of exceeding a threshold likelihood of collocation (see page 243, right column, lines 16-23, when lambda is greater than 1, a high likelihood of collocation exists and added to compound cluster. And see page 245, right column, "Simulation", 1st paragraph, if lambda is greater than zero then the n-gram is included in the compound list.) (e.g. From the cited portions a threshold of greater than one is chosen for lambda or the natural log of 1, which is greater than zero. If this occurs, then it is added to the compound list.)

7. Claims 4, 9, 16-18 and 27-29 is rejected under 35 U.S.C. 103(a) as being unpatentable over Su *et al.* in view of Frantzi *et al.* as applied to claims 1, 6, 13, and 24 above, and further in view of Manning (The MIT Press 1999).

As to claims 4, 9, 16-17, 27, and 28 Su *et al.* in view of Frantzi *et al.* teaches all of the limitations as in claims 1, 6, 13, and 24 above.

Furthermore, Su *et al.*, teaches where the likelihood ratio λ is computed by: $\lambda=(P(x\_|M_c)*P(M_c))/(P(x\_|M_{nc})*P(M_{nc}))$(see Su *et al.*, page 243, right column, line 9 (equation)) (e.g. It should be noted that the reference uses a different notation, but the same result and definitions are used, where the numerator is the n-gram produced by a compound result and the denominator is the result produced by a non-compound result. The formula can be changed to account for various distributions (Gaussian, Binomial).

However, Su *et al.* in view of Frantzi *et al.* do not specifically teach the likelihood ratio given by $\lambda=L(H_i)/L(H_c)$.

Manning shows the use of the likelihood ratio (see equation 5.10)(e.g. The equation in given in log form. The logs can be omitted to obtain the desired formula. The numerator is the independent hypothesis and the denominator is the dependence hypothesis.)

It would have been obvious to one of ordinary skilled in the art to have modified finding of compounds in a text corpus as taught by Su *et al.* and Frantzi *et al.* with the formula as taught by Manning. The motivation to modify the former

is for collocation discovery (see Manning, page 172, sect. 5.3.4, 3rd paragraph, lines 1-4).

As to claims 18 and 29, Su *et al.* in view of Frantzi *et al.* teaches all of the limitations as claim 13, 16, and 17 above.

Su *et al.* in view of Frantzi teach a system for identifying compounds through measure of association.

However, Su *et al.* in view of Frantzi do not specifically teach the representation of the independence and collocation hypothesis.

Manning does teach the explanations of these two types of hypothesis (see page 172, sect. 5.3.4, bullet items) (e.g. It should be noted that the independence hypothesis is given by hypothesis 1 and the dependence or collocation hypothesis by hypothesis 2. The w_2 and w_1 can be interpreted as the tokens since the reference deals with a text corpus).

It would have been obvious to one of ordinary skilled in the art to have modified the finding of compound words in a text corpus as taught by Su *et al.* and Frantzi *et* al with the inclusion of the two hypothesis as taught by Manning. The motivation to modify the former is for collocation discovery (see Manning, page 172, sect. 5.3.4, 3rd paragraph, lines 1-4). Further, the use of the formula presented by Manning would require an explanation of frequency for each type of hypothesis in order to find the likelihood ratio (definition of likelihood ratio).

Allowable Subject Matter

8. Claims 5, 10, 19, 25, and 30 are objected to as being dependent upon a rejected base claim, but would be allowable if rewritten in independent form including all of the limitations of the base claim and any intervening claims.

9. Claim 14 would be allowable if rewritten to overcome the rejection(s) under 35 U.S.C. 112, 2nd paragraph, set forth in this Office action and to include all of the limitations of the base claim and any intervening claims.

10. The following is a statement of reasons for the indication of allowable subject matter: none of the prior art references alone or in combination teaches or fairly suggests the limitations where "a limiter identifying a number of n-grams up to the upper limit based on number of occurrences" as seen in claims 14 and 25. Also, the limitations of "dividing the *n*-gram into *n*-1 pairings of segments... selecting the maximum likelihood of collocation of the pairings as $L(H_c)$" as seen in claims 5 and 10. Further, the limitations "$L(H_i)$ is computed ... in accordance with the formula:

$$\arg\max_{L(H_i)} \frac{L(t_1, t_2 formcompound)}{L(n-gramdoesnotformcompound)}$$" as seen in claims 19 and 30.

Conclusion

11. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure.

Smadja ("Retrieving Collocations from Text: Xtract") is cited to disclose retrieval and identification of collocations for textual corpora utilizing n-grams.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to Paras Shah whose telephone number is (571)270-1650. The examiner can normally be reached on MON.-THURS. 7:00a.m.-4:00p.m. EST.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Patrick Edouard can be reached on (571)272-7603. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

P.S.

02/04/2008